University of Michigan **University of Cape Town**

The Cape Area Panel Study:

A Very Short Introduction to the Integrated Waves 1-2-3-4-5 (2002-2009) Data

David Lam, Cally Ardington, Nicola Branson,
Anne Case, Murray Leibbrandt, Brendan Maughan-Brown,
Alicia Menendez, Jeremy Seekings and Meredith Sparks.

October 2012

Contents

1. Introduction

This document is intended as a tool for users of the CAPS Integrated Waves 1-2-3-4-5 data in their analyses by providing summary information on the structure of the datasets, including public identifiers, variable naming conventions and constructed variables. This document also provides information on how to correctly use sample weights and how to access and cite the data. For more indepth information on all the topics covered in this introduction please refer to the *Overview and Technical Documentation for Waves 1-2-3-4-5.*

The Cape Area Panel Study (CAPS) is a longitudinal study of the lives of youths and young adults in metropolitan Cape Town, South Africa. The first wave of the study collected interviews from about 4800 randomly selected young people age 14-22 in August-December, 2002. Wave 1 also collected information on all members of these young people's households, as well as a random sample of households that did not have members age 14-22. A third of the youth sample was re-interviewed in 2003 (Wave 2a) and the remaining two-thirds were re-visited in 2004 (Wave 2b). The full youth sample was then re-interviewed in both 2005 (Wave 3) and 2006 (Wave 4). Wave 5 – full face-to-face interviews – was conducted in 2009 with the sample comprising all respondents interviewed in any of Waves 2a, 3 or 4. Wave 5 continued in 2010 with short telephonic interviews with respondents who were not successfully interviewed during the 2009 fieldwork, or, if the respondent was unavailable, a proxy (someone who knows the respondent well). Wave 3 also includes interviews with approximately 2000 co-resident parents of young adults. Wave 4 also includes interviews with a sample of older adults (all individuals from the original 2002 households who were born on or before 1 January 1956) and all children born to the female young adults. Wave 5 also includes an HIV test administered among black African respondents. The study covers a wide range of outcomes, including schooling, employment, health, family formation, intergenerational support systems, and social and political attitudes and behavior.

CAPS began in 2002 as a collaborative project of the Population Studies Center in the Institute for Social Research at the University of Michigan and the Centre for Social Science Research at the University of Cape Town (UCT). Other units involved in subsequent waves include UCT's Southern African Labour and Development Research Unit, and the Research Program in Development Studies at Princeton University. Primary funding is provided by the National Institute of Child Health and Human Development of the U.S. National Institutes of Health (NIH). Additional funding has been provided by the Office of AIDS Research, the Fogarty International Center, the National Institute of Aging of NIH, the Health Economics & HIV/AIDS Research Division at the University of KwaZulu-Natal, the European Union (through the Microcon research partnership on the microfoundations of violent conflict, via the CSSR) and by grants from the Andrew W. Mellon Foundation to the University of Michigan and the University of Cape Town.

Section 2 lists the questionnaires included in each of the four waves. Section 3 contains a discussion of the variable naming conventions used in the CAPS data that both distinguish waves and modules as well as help identify equivalent variables across waves. Section 4 details the organisation of the public release datasets. Section 5 provides information on the sampling weights. Section 6 provides information on how to access the CAPS data as well as the acknowledgement and citation to use for CAPS.

2. CAPS Questionnaires

Table 1 below shows the questionnaires that were included in each of the four waves of CAPS. Please see the *Overview and Technical Documentation for Waves 1-2-3-4-5* for more details on the survey instruments.

Table 1: CAPS Questionnaires

	Wave 1 (2002)	Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)	Wave 5 (2009/10)
Young Adult Questionnaire	x	x	x	x	x	x
Young Adult Telephonic Questionnaire						x
Household Questionnaire	x			x	x	
Literacy and Numeracy Evaluation	x					
Parent Questionnaire				x		
Young Adult Proxy Questionnaire					x	x
Older Adult Questionnaire					x	
Child Questionnaire					x	

3. Variable naming conventions

Before detailing the organisation of the public release datasets it is necessary to briefly describe the variable naming conventions used in the CAPS data. Naming conventions are designed to simplify working with the CAPS data as a panel and include prefixes/suffixes which identify wave and questionnaire and renaming for consistency in the panel.

3.1. Original variables

All original (i.e. non-derived variables) CAPS variables begin with a prefix indicating the wave number and questionnaire type. Prefixes are added in order to avoid confusion between variables with the same name or number in different questionnaires and waves. For example, variable *w1h_a14* refers to question A14 in the Wave 1 (2002) Household questionnaire while *w4y_d1* refers to question D1 in the Wave 4 (2006) Young Adult questionnaire.

Variables that appear in multiple waves of CAPS have been renamed so that the name (after the prefix) is constant across waves. For example, D45 from Wave 1, F16 for Wave 2a, H8 from Wave 2b, D27 from Wave 3, D17 from Wave 4 and D18 from Wave 5 all ask the same question: Have you looked for work in the last month? These variables are renamed *w1y_lookwrk30*, *w2y_ lookwrk30*, *w3y_ lookwrk30*, *w4y_ lookwrk30, and w5y_ lookwrk30*. Variable labels for these variables indicate the original question number. These variable renames are inserted into all questionnaires under the original question number. All renamed variables are listed in the *CAPS Waves 1-2-3-4-5 Panel Variable Crosswalk* with both original question number and the rename specified.

Variables appearing in Waves 2a and 2b have the prefixes w2a_ and w2b_ respectively unless the variable appears in both Waves 2a and 2b and has been renamed, in which case w2y_ is used as the prefix.

3.2. Derived variables

3.2.1. Panel variables

Panel variables for each year from 2002 to 2009 are created using data from multiple waves. Information for a particular variable may be drawn from different waves for different respondents. For example information about whether the respondent was in school in 2002 could possibly come from any of the following questions: *w1y_b6*, *w1y_c6*, *w2a_e10*, *w2b_e9*, *w3y_c3*, *w3y_c5*, *w3y_c35* or *w4y_c1*. Panel variables have a year suffix (*_year*) instead of a prefix. For instance, *insch_02* is a panel variable which could be created from information from any of the waves and indicates whether the respondent was in school in 2002.

The creation of derived variables often requires choices and reflects assumptions. Some respondents reported inconsistent data in successive interviews, for example reporting that they had completed different levels of schooling in the specified year. Sometimes the grade given in the more recent wave was lower than the original grade reported. In the variables created to indicate highest level of schooling the choice was made to use the grade reported in the earliest wave. The do files used to create the derived variables are available online (see Section 6.4) and should be reviewed to understand how the variables were created.

3.2.2. Updating variables

Updating variables indicate whether the respondent had experienced an event by a particular wave. Once the respondent has experienced the event, the updating variable remains unchanged for all following waves. Updating variables have a suffix indicating the wave number and questionnaire type. For example *hadsex_w3y* refers to whether the respondent has had sexual intercourse sometime before the wave 3 interview.

3.2.3. Constructed variables

We have constructed a number of variables that we think will be helpful to researchers. These variables begin with the appropriate prefix indicating wave and questionnaire type, but the variable name is not based on the questionnaire. For example *w4y_wrknow* indicates whether the respondent was working at the time of the Wave 4 interview and is created from *w4y_d9_j1* to *w4y_d9_j6*. Variable labels for constructed variables indicate the original questionnaire variables that were used to construct them.

All derived variables are listed in the *CAPS Waves 1-2-3-4-5 Panel Variable Crosswalk* with both original and derived variable names. Table 2 below shows all prefixes/suffixes used in the data for CAPS Waves 1, 2, 3, 4 and 5.

Table 2: Prefixes and suffixes

Prefix/ Suffix	Wave(s)	Module
w1h_	1	Household
w1y_	1	Young Adult
w1y_lne	1	Young Adult Literacy and Numeracy Evaluation
w2a_	2a	Young Adult
w2b_	2b	Young Adult
w2y_	2	Young Adult (renamed variables which are in both the w2a and w2b questionnaires)
w2h_	2	Household roster (from young adult questionnaire)
w3y_	3	Young Adult
w3h_	3	Household
w3p_	3	Parent
w4y_	4	Young Adult
w4x_	4	Young Adult Proxy
w4h_	4	Household
w4o_	4	Older Adult
w4c_	4	Child
w5y_	5	Young Adult
w5t_	5	Young Adult Telephonic
w5x_	5	Young Adult Proxy
w5h_	5	Household
_02	panel (1-2-3-4)	Young Adult 2002 information (Non wave-specific prefix variables only)
_03	panel (1-2-3-4)	Young Adult 2003 information (Non wave-specific prefix variables only)
_04	panel (2-3-4)	Young Adult 2004 information (Non wave-specific prefix variables only)
_05	panel (3-4-5)	Young Adult 2005 information (Non wave-specific prefix variables only)
_06	panel (4-5)	Young Adult 2006 information (Non wave-specific prefix variables only)
_07	panel (5)	Young Adult 2007 information (Non wave-specific prefix variables only)
_08	panel (5)	Young Adult 2008 information (Non wave-specific prefix variables only)
_09	panel (5)	Young Adult 2009 information (Non wave-specific prefix variables only)
_w1y	updating (1)	Young Adult up until Wave 1
_w2y	updating (1-2)	Young Adult up until Wave 2
_w3y	updating (1-2-3)	Young Adult up until Wave 3
_w4y	updating (1-2-3-4)	Young Adult up until Wave 4
_w5y	updating (1-2-3-4-5)	Young Adult up until Wave 5
sp01_	--	Community data – subplace level
mp01_	--	Community data – mainplace level

4. Organization of the public release datasets

The 2012 release of the CAPS data includes Waves 1, 2, 3, 4 and 5, covering 2002-2009. Some questions and modules are not included in the public release of CAPS in order to protect the confidentiality of our respondents. For some of these variables, the original variable is replaced with a coded version of the variable, which does not risk a breach of confidentiality.

The types of variables that are not included are names (of respondents, household members, employers, children/partners of young adults); addresses, phone numbers and all other contact details; names of schools attended by respondents; detailed descriptions of jobs; day of birth (month and year are included); and all identifiers for individuals, households, and neighborhoods are scrambled and have no correspondence to codes in the South Africa census. Names of schools are replaced with codes (which have no meaning beyond CAPS) and descriptions of jobs are replaced with Standand Occupation and Industry (SOC/SIC) codes. Additionally, the Wave 2b (2004) School choice module (questions E.42-E.58) and town and school names from the Wave 3 Residential and Schooling History (B.7 B.9 B.10) are not included in the public release data.

All of the data cleaning and data analysis of CAPS data by project staff has been done using Stata 11, SE. The data is available in the following formats: Stata 11 SE, Stata 8 SE, SPSS, SAS and ASCII with dictionary files. See section 6 of this document for details of how to download the data.

The CAPS Stata data files included in the Waves 1-2-3-4-5 2012 release are summarized in Table 3. The name of each dataset contains information about the wave(s), questionnaire (e.g. household or young adult), type and version of the data. The "*v*" represents a version number that appears in each file name, with the first two digits indicating the year and last two digits indicating the month. For example, the capsw2345.y.cal.wide.v1210.dta file is the Wave 2-3-4-5 young adult (questionnaire) calendar in wide form (type) dated October 2012 (version).

A short discussion highlighting some of the important features of each dataset follows Table 3. Please see the *Overview and Technical Documentation for Waves 1-2-3-4-5* for more details on these datasets.

Table 3: Public release datasets

File name	**Description**	**Number of records**	**Unique identifier**	**File structure**
household:				
capsw1.h.roster.*v*.dta	Wave 1 household roster file	22629	personid	One record per household member
capsw2.h.roster.*v*.dta	Wave 2 resident members household roster file	15157	personid w2h_hhid	One record per household member, except for 1 duplicate
capsw3.h.roster.*v*.dta	Wave 3 resident members household roster file	12994	personid	One record per household member
capsw4.h.roster.*v*.dta	Wave 4 resident members household roster file	15648	personid w4h_hhid	One record per household member, except for 24 duplicates
capsw5.h.roster.*v*.dta	Wave 5 resident members household roster file	12350	personid w5h_hhid	One record per household member
capsw1.h.*v*.dta	Wave 1 household level file	5255	hhid	One record per household
capsw3.h.*v*.dta	Wave 3 household level file	2549	w3h_hhid	One record per household
capsw4.h.*v*.dta	Wave 4 household level file	3312	w4h_hhid	One record per household
capsw5.h.*v*.dta	Wave 5 household level file	2313	w5h_hhid	One record per household
capsw2.h.nr.*v*.dta	Wave 2 non-resident members household file	1572	personid w2h_hhid	Duplicate records on personid
capsw3.h.nr.*v*.dta	Wave 3 non-resident members household file	2404	personid w3h_hhid	Duplicate records on personid
capsw4.h.nr.*v*.dta	Wave 4 non-resident members household file	3796	personid w4h_hhid	Duplicate records on personid
young adult:				
capsw12345.y.*v*.dta	Merged wave 1-2-3-4 young adult individual file	5291	personid	One record per young adult
capsw12345.y.derived.*v*.dta	File of created, panel, updating and job table variables	4752	personid	One record per young adult
capsw1.y.lne.*v*.dta	Wave 1 complete literacy/numeracy evaluation	4742	personid	One record per young adult

Table 3 (continued)

File name	**Description**	**Number of records**	**Unique identifier**	**File structure**
Calendar:				
capsw1.y.cal.wide.*v*.dta	Wave 1 young adult life history calendar in wide form	4752	personid	One record per young adult
capsw1.y.cal.long.*v*.dta	Wave 1 young adult life history calendar in long form	89556	personid w1y_calage	One record per year of each young adult's life
capsw2345.y.cal.wide.*v*.dta	Merged wave 2-3-4-5 young adult monthly calendar in wide form	4752	personid	One record per young adult
capsw2345.y.cal.long.*v*.dta	Merged wave 2-3-4-5 young adult monthly calendar in long form	455540	personid capsmth	One record per month for each young adult from August 2002 to most recent interview date
Other files:				
capsw1.h.nrc.*v*.dta	Wave 1 file of non-resident biological children aged 0-22	2635	personid	One record per non-resident child
capsw3.p1.*v*.dta	Wave 3 parent file	1967	personid	One record per parent
capsw3.p2.*v*.dta	Wave 3 parent file	2707	Personid -once drop duplicates	One record per young adult, 4 duplicates
capsw4.c.*v*.dta	Wave 4 child file	922	personid	One record per child
capsw4.o.*v*.dta	Wave 4 older adult file	3564	personid	One record per older adult
capsw4.x.*v*.dta	Wave 4 proxy file	285	personid	One record per young adult
capsw5.t.*v*.dta	Wave 5 telephonic file	262	personid	One record per young adult
capsw5.x.*v*.dta	Wave 5 proxy file	84	personid	One record per young adult
capsw5.HIV.v.dta	Wave 5 HIV file	1248	personid	One record per young adult
capsw5.h.duplicates.v.dta	Wave 5 duplicates file	1756	personid w5h_hhid	One record per household member
capsw5.h.triplicates.v.dta	Wave 5 triplicates file	167	personid w5h_hhid	One record per household member
capsw5.y.discarded.v.dta	Wave 5 YA suspected fraudulent	234	personid	One record per young adult
capsw5.h.roster.discarded.v.dta	Wave 5 household rosters suspected fraudulent	921	personid w5h_hhid	One record per household member
Additional resources:				
capsw1.h.community.*v*.dta	Subplace and mainplace variables from 2001 South Africa census file matched to CAPS neighborhoods	405	cluster	One record per neighbourhood
capsw1234.y.schoollevel.*v*.dta	Data from the 2000 South Africa School Register of Needs matched to CAPS schools	2983	code	One record per school

4.1. Household files

4.1.1. Wave 1 household files

These files contain information from all CAPS households visited in Wave 1, including households without young adults. The roster file (capsw1.h.roster.*v*) contains all the information from module A (roster of members) of the household questionnaire and is organized as one record per household member, with the number of records varying from household to household. The roster format of the household data facilitates the generation of household-level variables, and can also be used to assign characteristics of other household members to young adults within the household. Variables from modules B, C and D of the household questionnaire are common to all household members. These variables are found in the capsw1.h.*v* file which has only one record per household.

4.1.2. Wave 2 household files

Wave 2 did not include a separate household questionnaire but a household roster was attached to each individual young adult questionnaire. This information is contained in capsw2.h.roster.*v*. In households with multiple young adults, one representative young adult household module is included in the capsw2.h.roster.*v*. data. Thus in some cases a young adult may not complete the individual questionnaire but have information in the household file. This information comes from a successfully completed co-resident young adult in their household. *w2h_finalresult* indicates whether the respondent has household information, regardless of whether they completed the individual questionnaire or not.

No household level questions were asked. In order to assist with the identification of household members across waves, all individuals co-residing with the young adult in Wave 1 were pre-printed on the household roster section of the young adult questionnaire. The fieldworker then recorded who was still co-resident with the young adult and added any additional household members to the roster. Individuals who were no longer co-resident with the young adult are not included in the capsw2.h.roster.*v* file. These individuals and the reason given for why they were no longer living with the young adult can be found in the non-resident household file, capsw2.h.nr.*v*.

4.1.3. Waves 3 and 4 household files

Waves 3 and 4 included separate household questionnaires similar to that used in Wave 1. There are three household datasets from each of Wave 3 and 4. The roster files (capsw3.h.roster.*v* and capsw4.h.roster.*v*) include all information from section B and are organized as one record per household member. Information from sections C to G is found in capsw3.h.*v* and capsw4.h.*v* for Waves 3 and 4 respectively. As with Wave 2, all individuals co-residing with a young adult in the previous wave were pre-printed on the household roster. Individuals who were no longer living with the young adult in Wave 3 and 4 can be found in capsw3.h.nr.*v* and capsw4.h.nr.*v* respectively.

4.1.4. Waves 5 household files

Wave 5 included household questions in the young adult questionnaire. A set of screeening questions (page 4) were used to determine whether the young adult completed the household roster (pages 6 and 7). As with Waves 2, 3 and 4 all individuals co-residing with a young adult in the previous wave were pre-printed on the household roster. The roster file (capsw5.h.roster.v) includes all inforamtion from the household roster and is organized as one record per household member. Despite the household roster screening questions, 147 household rosters were completed by two co-resident young adults and 6 household rosters by three co-resident young adults. These duplicate and triplicate household rosters have been combined into one roster – see the *Overview and Technical Documentation for Waves 1-2-3-4-5* for more details. The original data for the duplicate and triplicate rosters are contained in separate data sets (capsw5.h.duplicates and capsw5.h.triplicates) – these are available on request.

All young adults were asked the household level questions in the Wave 5 questionnaire. There are therefore two or three data points for households containing two or three co-resident young adults. These household level questions are found in module B and at the end of module D. The household level data (capsw5.h.v) was created using the data point from the oldest young adult who answered each question ("don't know" or "refused" answers not included). The assumption is that the oldest respondent was most knowledge about the household. Users who wish to create household level data based on different assumptions can find the original data in the young adult data set.

4.1.5. Wave 5 discarded fraudulent household roster data

The capsw5.h.roster.discarded.v.dta file contains the data collected in 2009 that were suspected of being collected through fraudulent fieldwork and have been dropped from the household roster dataset (for more details see capsw12345.overview&technical.v.doc, Section 4.5.1). The variable "w5h_fraud_cat" indicates the category of fraud detected:

1. Completely fraudulent: the entire interview appears fabricated.
2. Short telephone interview: some basic demographic information collected via telephone.
3. Fraud suspected, but not verified: interviews conducted that could not be validated because the respondent was not reachable. These interviews were conducted after other fieldwork that was proven fraudulent.

4.1.6. Merging household rosters across waves

The household level files are presented separately for each wave since the CAPS was not designed as a household panel. Some care was taken to match individual household members across waves but this was not the focus of the study so there may be some error in the matching of people who are not in the young adult or older adult samples. For example, when household members move out of and then later back into a household they may erroneously be assigned a second *personid*. That said, in the majority of cases a household member with the same *personid* over time is the same person and users can merge the household roster files across waves using the *personid* variable.

4.2. Young Adult files

4.2.1. Waves 1-2-3-4-5 Young Adult data

The capsw12345.y.*v* file contains all of the information collected in the Waves 1, 2a, 2b, 3, 4 and 5 young adult (YA) questionnaires. The file contains one record per young adult. Young adults who were selected into the young adult sample but did not complete questionnaires are included in the file even though all or most of their records are missing. The 4,752 young adults that form the entire young adult panel can be identified using the variable *panelya*[1], which has a value of 1 for panel members. This variable is found in all household roster datasets.

Young adults who did not complete one of the subsequent waves of the panel are included in the file with missing data for the wave(s) they did not complete. The code indicating why the questionnaire was not completed (refused, not available, etc.) is included in the **finalresult* variables.

Given the sample design, many households have more than one YA in the YA sample. The variable *yanum* gives the number (1, 2, or 3) of the YA in the household, including those who did not complete the Wave 1 questionnaire and are therefore not part of the young adult panel.

In the Wave 3 young adult questionnaire questions on puberty (E9a and E9b) and questions on sexual debut (E10 to E15) were omitted from the questionnaire if the respondent had indicated that they had reached puberty or had sexual intercourse in waves 1 or 2. In the Wave 5 young adult questionnaire questions that were redundant or not applicable to respondents were marked with an "x" and not asked during the interview. Please see *Overview and Technical Documentation for Waves 1-2-3-4-5* for further details on questions that were greyed out or pre-loaded with information from previous waves.

The vast majority of fieldwork for Wave 1 was completed in 2002. There were however a few interviews conducted in early 2003. Similarly a small number of Wave 3, 4 and 5 interviews took place in early 2006, 2007 and 2010 respectively. Analysts should be careful to check the interview date as the interpretation of some variables may differ depending on when the interview was conducted.

4.2.2. Waves 1-2-3-4-5 Young Adult derived data

All derived (panel, updating or constructed) variables can be found in capsw12345.y.derived.*v*. This file can be merged with capsw12345.y.*v* using *personid*. The different kinds of panel variables were briefly discussed in section 3 above. The numbering of jobs within each wave and across the panel is somewhat complicated and is therefore explained in some detail in this section.

In Waves 3 and 4, any job in which a respondent was working at the time of the previous interview is pre-loaded into the YA's questionnaire. For example information on any current jobs from Wave 2 is pre-loaded into the Wave 3 questionnaire. The aim is for subsequent interviews to update and complete the information for this job, including the reason why the respondent stopped working at the job if applicable. This means that there is potentially information from multiple waves on the same job. For example, if a respondent was working in the same job from January 2003 to January 2006 and was interviewed in every wave, we would have asked how much they earned in this job in waves 2, 3 and 4. Within each wave jobs, including those that were pre-loaded, are numbered starting from one. Wave specific jobs can be identified by the suffix _#. For example w4y_d11_1 is the variable for the earnings for job 1 in Wave 4. All the original job variables are included in the capsw12345.y.*v* dataset.

In the capsw12345.y.derived.*v* file, a w234 job table series is created collating the information from each of Waves 2, 3 and 4 for every job that the respondent reported since the Wave 1 interview. In order to do this we

[1] YA's who form part of the panel can alternatively be identified by *w1y_finalresult* taking on the value 1.

need a job numbering system that identifies the same job across different waves. The way in which this is achieved is best explained by an example. Assume our respondent is interviewed in each of waves 2a, 3 and 4. In Wave 2a they reported two jobs since their Wave 1 interview. At the time of the Wave 2a interview they were no longer working in the first job but were still working in the second job. In Wave 3 the second job from Wave 2a would have been pre-printed on their questionnaire. They reported that they stopped working in this job and that they had since had another job but that job had also ended. In Wave 4 no jobs would have been pre-printed on the questionnaire as the respondent was not currently working when last seen in Wave 3. The respondent reported a new and ongoing job in wave 4. The numbering of the jobs from this example within each wave and in the w234 job table is shown in Table 4 below. Job 2 in Wave 2a and job 1 in Wave 3 correspond to job 2 in the w234 job table. Likewise job 1 in Wave 4 is job 4 in the w234 job table. The series of w#y_jobnum# variables identify the wave specific job numbers for each job in the w234 job table. For example, w3y_jobnum2 gives the wave 3 job number of job 2 in the w234 job table. In the example shown in Table 4 below, w3y_jobnum2=1.

Table 4: Example of job numbers

W234 Job Table	Wave 2a	Wave 3	Wave 4
1	1		
2	2	1	
3		2	
4			1

Additional information on work is available from wave 5, but these data have not yet been incorporated into these composite job variables.

4.2.3. Wave 1 Literacy and Numeracy Evaluation

The capsw1.y.lne.*v* file contains the complete information from the young adult literacy and numeracy evaluation (LNE). This includes the answers given to all 45 questions, along with the number of correct answers to the literacy questions, the number of correct answers to the numeracy questions, the combined total correct, standardized scores and the language in which the evaluation was taken (English or Afrikaans).

4.3. Calendar data

All calendar files are available in long and wide format. The wide format has one record per young adult while the long format has one record per month for each young adult.

4.3.1. Wave 1 calendar data

In the wide format of the calendar data (capsw1.y.cal.wide.*v*) each age is identified by the suffix at the end of the variable name. For example, if *w1y_b1_12*=1, this means that the YA moved household in the year that they were 12 years old. In the long form of the calendar data (capsw1.y.cal.long.*v*) each age if identified by the variable *w1y_calage*. For example, if *w1y_b2a*=1 when *w1y_calage*=5, this means that the YA lived with her mother at least half of the year when she was 5 years old.

All researchers working with calendar data should carefully read the discussion about the two different time perspectives used in the calendar presented in the section on the Wave 1 Life History Calendar Questionnaire in the *Cape Area Panel Study: Overview and Technical Documentation for Waves 1-2-3-4-5*. The interpretation of *w1y_calage* in the long form or the age suffix in the wide form differs for variables on page 1 and page 2 of the calendar. For variables on page 1 of the calendar, which deal with household living arrangements and relationships, *w1y_calage* (long form) or the age suffix (wide form) should be thought of simply as the age at which these events took place. For example in the long form, if *w1y_b2b*=2 when *w1y_calage*=7, this means that the YA did not live with her father at least half of the year when she was 7 years old. On page 2 of the calendar, which covers school, work, and pregnancy, *w1y_calage* refers to the age of the YA at the beginning of a calendar year. For example, if *w1y_b6*=1 when *w1y_calage*=8, this means that the YA was enrolled in school in the year in which she was age 8 at the beginning of the year. If the YA had a birthday early in the year, the YA was actually age 9 for most of that school year. If the YA had a birthday late in the year, the YA was age 8 for most of the year. Researchers interested in looking at something like school attendance or work at a given age will want to keep these issues in mind.

4.3.2. Wave 2-3-4-5 Young adult monthly calendar data

All variables in the wide version (capw2345.y.cal.wide.*v*) of the Waves 2-3-4-5 young adult monthly calendar are named using the following "CAPS-months" naming conventions as a suffix. August 2002 is the first month of the monthly calendar data, and is month "8" in CAPS-months. For example, the variable *w2a_f1_m9* in the wide monthly calendar data contains information on school enrollment for September 2002 from wave 2a. In the long version of the monthly calendar data (capsw2345.y.cal.long.*v*), the values of variable *calmonth* for each row of data are recorded in CAPS-months and the variable is value-labeled with the corresponding real calendar month and year.

In addition to the calendar variables from each of the individual waves, the calendar files include variables that combine monthly calendar data from Waves 2a, 2b, 3, 4 and 5, producing a seamless record from August 2002 until the month of each young adult's most recent interview. The questionnaires were designed to fill in the calendar from the date of the last interview. As a result, a respondent who was not seen in Waves 2 or 3 but was re-interviewed in Wave 4 will have complete calendar information from August 2002. It is possible that data, for instance, for February 2003 may have been collected in Wave 2a, 2b, 3 or 4, depending on when the young adult was successfully re-contacted and interviewed.

The job numbers corresponding to the *calwork** variables match the job numbers in the w234 job table.

In the wide monthly calendar data, respondents are coded "97" for all months after the month of their last successful interview. In the long data, the months after last successful interview have been deleted.

4.4. Other individual wave data files

4.4.1. Wave 1 Non-resident biological children file

The capsw1.h.nrc.*v* file contains the information collected in the section of the Wave 1 household questionnaire dealing with children of household members who are age 0 to 22 and are not resident in the household. This information is collected in the two-page roster that contains questions A.36 through A.53. The file is organized as one record per non-resident child, with the number of records varying from household to household. Variable names begin with *w1h_*, since these variables are all taken from the household questionnaire. The variables *w1h_a37* and *w1h_a38* give the line numbers for the non-resident child's mother and father, respectively, in the household roster. By definition at least one parent will reside in the household and will have a non-missing line number on one of these variables.

4.4.2 Wave 3 Parent files

The Wave 3 Parent data is available in two formats, representing two different perspectives with which to analyze the data. All variables in both datasets from the parent questionnaire have the prefix *w3p_*.

The file w3.p1.*v*.dta (format 1) is organized as one record per parent and can be merged with the household roster files on the variable *personid*.

In the file w3.p2.*v*.dta (format 2), the parent data has been reshaped to one record per young adult, so that the YA-specific responses in the parent data are matched to the corresponding young adult in the young adult data when merged on the variable *personid*. Variables that are common to all household members, such as information regarding the "selected child" and the parent-respondent, are attached to every young adult in the this file.

4.4.3. Wave 4 child file

The capsw4.c.*v* file contains all the information from the Wave 4 Child questionnaire. In Wave 4 we attempted to interview the primary caregiver and take anthropometric measures for all children born to female young adults who were successfully re-interviewed in Wave 4. This includes children who were not co-residing with their mothers at the time of the Wave 4 interview.

4.4.4. Wave 4 older adult file

The capsw4.o.*v* file contains all the information from the Wave 4 Older Adult questionnaire. The older adult sample consists of all individuals resident in original Wave 1 households who were born on or before 1 January 1956. This includes Wave 1 households without young adults.

4.4.5. Wave 4 young adult proxy

The capsw4.x.*v* file contains all the information from the Wave 4 Young Adult Proxy questionnaire. In cases where a young adult was not available, interviewers were instructed to conduct a proxy interview if they found an individual who was knowledgeable about the young adult.

4.4.6. Wave 5 young adult telephonic interviews

The capsw5.t.*v* file contains all the information from the Wave 5 Young Adult Telephonic questionnaire. Between February and October of 2010 UCT attempted to update core CAPS data for respondents who were not successfully interviewed during the 2009 fieldwork. In addition, these core data were collected for respondents whose initial interview in 2009 was suspected fraudulent (see capsw12345.overview&technical.v.doc, Section 4.5.1). Respondents were contacted via telephone and asked about their current contact details, household, education, employment and some demographic information.

4.4.7. Wave 5 proxy interviews

The capsw5.x.*v* file contains all the information from the Wave 5 Proxy questionnaire. In cases where we were unable to interview the Young Adult, face-to-face or telephonically, we sought to collect basic information via telephone from some one who knows the Young Adult well (usually a close family member).

4.4.8. Wave 5 HIV test results

The capsw5.HIV.*v* file contains the HIV test results. Access to these data is restricted. Individuals who wish to use this data must submit a research proposal to Professor Jeremy Seekings (jeremy.seekings@uct.ac.za) or Professor David Lam (davidl@isr.umich.edu).

4.4.9. Wave 5 discarded fraudulent young adult data

The capsw5.y.discarded.v.dta file contains the data collected in 2009 that were suspected of being collected through fraudulent fieldwork and have been dropped from the young adult dataset (for more details see capsw12345.overview&technical.v.doc, Section 4.5.1). The variable “w5y_fraud_cat” indicates the category of fraud detected:

1. Completely fraudulent: the entire interview appears fabricated.
2. Short telephone interview: some basic demographic information collected via telephone.
3. Fraud suspected, but not verified: interviews conducted that could not be validated because the respondent was not reachable. These interviews were conducted after other fieldwork that was proven fraudulent.

4.5. Additional resources

4.5.1. Community data

The data used for the construction of the CAPS Wave 1 community level data file (capsw1.h.community.*v*) are based on the 2001 South Africa Census tabulated data at the subplace and mainplace level. These data were accessed using the SuperCross software supplied with the tabulated data by Statistics South Africa. See the *Overview and Technical Documentation for Waves 1-2-3-4-5* for more details.

4.5.2. School data

The file capsw1234.y.schoollevel.*v*, contains school level data on all the schools attended by CAPS respondents. The School Register of Needs (2000) is a survey of nearly 30,000 schools and other public educational institutions through the secondary level throughout all nine provinces of South Africa. The survey instrument asks questions relating to school infrastructure, services, available equipment and resources. There are multiple questions (e.g. *w1y_c11cd* , *w4y_c5_02cd*, *w3y_b9_00cd schcd_02* etc) in the capsw12345.y.*v*.dta and capsw12345.y.derived.*v*.dta which identify schools. Users should pick the variable they are interested in, e.g. *schcd_02*, rename this variable *code* and then merge the YA data with the schoollevel file on *code*.

4.5.3. Crime data

A separate do file entitled “capsw3&5.crime.rates.by.magisterial.district.do” links SAPS crime statistics by police precinct to the CAPS data at the magisterial district level for Waves 3 and 5. The SAPS crime categories are: all other theft not mentioned elsewhere, arson, assault with the intent to do grievous bodily harm, attempted murder, burglary at business premises, burglary at residential premises, carjacking, commercial crime, common assault, common robbery, crimen injuria, culpable homicide, driving under the influence of alcohol or drugs, drug-related crime, illegal possession of firearms and ammunition, kidnapping, malicious damage to property, murder, neglect and ill treatment of children, public violence, robbery at business premises, robbery at residential premises, robbery with aggravating circumstances, shoplifting, theft of motor vehicle and motorcycle, theft out of or from motor vehicle and sexual crimes. The crime category truckjacking was excluded due to insufficient observations. The SAPS crime statistics were obtained from http://www.saps.gov.za/statistics/reports/crimestats/2009/provinces/w_cape/western_cape.htm. The magisterial districts are: Bellville Cape, Goodwood, Kuilsrivier, Malmesbury, Mitchellsplain, Simonstown, Somerset West, Strand and Wynberg. The SAPS crime statistics for Bellville were created by taking a simple average of the crime statistics for the Bellville and Bellville South police precincts for each crime category.

4.6. Updates to Version 0810

All v0810 public release data are available online with the new suffix v1210. In the majority of cases, apart from the new name, these data remain unchanged. The following section outlines the changes that have been made to specific datasets:

capsw1.h.v
The derived household income measures have been removed and replaced with new measures (see Section 11 of the *Overview and Technical Documentation for Waves 1-2-3-4-5*).

capsw3.h.v
Several of the household expenditure variables were mistakenly left out of the 0810 version of the data and are now included. These variables are w3h_d11-22(a)(b)&(c).
In addition, the derived household income measures have been removed and replaced with new measures (see Section 11 of the *Overview and Technical Documentation for Waves 1-2-3-4-5*).

capsw3.h.roster.v
The disability grant and child support grant variable labels were mixed up, i.e. they labeled each others data. These data now have the correct labels.

capsw12345.y.derived.v

1. The do file to create these data constructs a package that identifies the years with the most complete school information and then uses this package to create the school variables. In the latest version the definition of the package has changed. Previously the first wave with the most complete information on insch, schrslt and edlvlcur was used. In the new version the first wave with the most complete information on insch, schrslt, edlvlcur and schcd was used. This resulted in two types of changes:
   a. When there are inconsistencies between waves and the choice of package changed
   b. In some cases a different wave package was used and this resulted in additional information or missing information e.g. If in wave 1 have info on insch, edlvlcur and schcd and in wave 3 have info on insch edlvlcur and schrslt then in the previous file would have used wave 3 and hence have non missing info for those three vars, while in the new file you would use wave 1 and hence have non-missing information for insch edlvlcur and schcd but schrslt would be missing.
2. The inclusion of wave 5 - this impacts _06 variables in particular
3. Other changes - errors were found in the initial do file. These have been fixed, see comments marked "***NB:" in the do file *"CAPSW12345.PANELVARS"*

4.7. Important Wave 5 Data Decisions

Data users who use the Wave 5 respondent telephonic (capsw5.t.) or proxy (capsw5.x.) data should consider the time variant nature of these variables as variables were collected 8-12 months after the 2009 fieldwork. Should these data be used and weighted the appropriate weight should be applied (see Section 6 of the *Overview and Technical Documentation for Waves 1-2-3-4-5*).

Regarding the discarded fraudulent data, users may wish include some of these data in analysis. The fraud investigation (see capsw12345.overview&technical.v.doc, Section 4.5.1) discovered that in many cases the basic demographic data were collected from respondents via a short telephone interview and then the rest of the questionnaire fabricated. For example, the names of household members in the household roster, education completed, employment status and job descriptions were often valid. These potentially useful data are identified by the variables "w5y_fraud_cat" & "w5h_fraud_cat" (values "2" & "3") in the discarded datasets.

5. Sample design and weights

The CAPS household sample was drawn through a two-stage process. First, the 'enumeration areas' (EAs) used for the 1996 Population Census were divided into three strata according to whether the population of each was predominantly African, predominantly coloured or predominantly white. A sample of primary sampling units (PSUs) was selected within each stratum with probability proportional to size. Within each PSU a sample of 25 screener households was drawn. The *Overview and Technical Documentation for Waves 1-2-3-4-5* provides a more detailed discussion of the sampling design. Data users should take the stratification and clustering into account for all analyses. Strata and PSUs are identified by the *majpop* and *cluster* variables respectively.

The public release data include sample weights that should be used to adjust for the sample design and wave 1 non-response summarized in the *Overview and Technical Documentation for Waves 1-2-3-4-5*. Including the sample weights in analyses enables the following inference:
1. The Wave 1 household sample is representative of households in metropolitan Cape Town at the time of the survey, including both households with and without young adult residents.
2. The young adult sample is representative of the non-institutionalized population aged 14-22 in metropolitan Cape Town at the time of the survey.
Three sample weights are included in the data, each one dealing with specific issues.

Table 5: CAPS weights

Variable name	Adjustment	Applicability
weightsd	Adjusts for sample design	Relevant to all files
weighthr	*weightsd* + wave 1 household non-response	Relevant to all files
weightyr	*weighthr*+ wave 1 young adult non-response	Relevant to young adult file

See the *Overview and Technical Documentation for Waves 1-2-3-4-5* for information on weights that adjust for attrition between waves of the panel.

6. Accessing and citing the data

6.1. Accessing the data

The CAPS Waves 1-2-3-4-5 data can be accessed over the internet. Access is through the CAPS website:
http://www.caps.uct.ac.za
or through DataFirst at the University of Cape Town:
http://www.datafirst.uct.ac.za/

Users will be required to register before accessing the actual data.

6.2 Acknowledging and citing the data

Papers using the CAPS Waves 1-2-3-4-5 data should include the following acknowledgement:

The Cape Area Panel Study Waves 1-2-3 were collected between 2002 and 2005 by the University of Cape Town and the University of Michigan, with funding provided by the US National Institute for Child Health and Human Development and the Andrew W. Mellon Foundation. Wave 4 was collected in 2006 by the University of Cape Town, University of Michigan and Princeton University. Major funding for Wave 4 was provided by the National Institute on Aging through a grant to Princeton University, in addition to funding provided by NICHD through the University of Michigan. Wave 5 was collected in 2009 by the University of Cape Town. Major funding for Wave 5 was provided by the Health Economics & HIV/AIDS Research Division (HEARD) at the University of KwaZulu-Natal, with additional funding from the Andrew W. Mellon Foundation (through the CSSR at UCT), the European Union (through the Microcon research partnership on the microfoundations of violent conflict, via the CSSR) and the NICHD (through the University of Michigan).

We recommend the following citation for papers using CAPS Waves 1-2-3-4-5:

David Lam, Cally Ardington, Nicola Branson, Anne Case, Murray Leibbrandt, Brendan Maughan-Brown, Alicia Menendez, Jeremy Seekings and Meredith Sparks. *The Cape Area Panel Study: A Very Short Introduction to the Integrated Waves 1-2-3-4-5 Data*. The University of Cape Town, October 2010.

6.3 Keeping our record and website up to date

Copies of any seminar/conference papers or published work should be sent to: Lynn Woolfrey, Data First, Centre for Social Science Research, University of Cape Town, Private Bag, Rondebosch, Cape Town 7701, South Africa; email lynn.woolfrey@uct.ac.za – or to caps@umich.edu.

6.4 Documentation

All documentation and survey instruments are available on the CAPS and DataFirst websites.Documentation for CAPS Waves 1-2-3-4-5 includes:

A Very Short Introduction to the CAPS Integrated Waves 1-2-3-4-5 Data: This brief document is intended to familiarize analysts with the organization of the public release data sets. It is an essential read before using the data.

The Cape Area Panel Study: Overview and Technical Documentation for Waves 1-2-3-4-5 Introduces the major motivations for the CAPS project as a whole, the project team and sponsors, details of the original sample design, as well as describing fieldwork, training, the survey instruments, and response rates for each of Waves 1, 2, 3, 4 and 5.

CAPS Waves 1-2-3-4-5 Panel variables crosswalk: Matches variables which are repeated across the panel in merged datasets, details derived 'panel' and 'updating' variables.

Do files: CAPS users can review the code used to create the derived variables in a set of Stata "do" files. Further information is included in the *Overview and Technical Documentation for Waves 1-2-3-4-5.*

6.5 Enquiries

Enquiries about the data-set should be sent to caps@umich.edu. Queries about the weights should be sent to Professor Martin Wittenberg at DataFirst, UCT (martin.wittenberg@uct.ac.za; see www.datafirst.uct.ac.za). See the CAPS website for more contact information: http://www.caps.uct.ac.za/contact.html.